IDENTIFICATION OF PROGNOSTIC BASELINE VARIABLES TO BE USED IN THE ANALYSIS OF ALZHEIMER'S DISEASE AND MILD COGNITIVE IMPAIRMENT CLINICAL TRIALS

by

Melody Dehghan

A thesis submitted to The Johns Hopkins University

in conformity with the requirements for the degree of

Master of Science

Baltimore, Maryland

May, 2021

Abstract

Adjusting for baseline variables that are prognostic for the outcome can improve precision of estimated marginal treatment effects in randomized trials, thus improving power for a fixed sample size or reducing the required sample size to achieve a desired power.

We compare the statistical properties of the estimated marginal treatment effect derived from the analysis of covariance and unadjusted estimators when we pre- vs. post-select baseline variables using several variable selection procedures within the context of an on-going Alzheimer's Disease (AD). Our simulation studies mimic the on-going HOPE4MCI trial, whose goal is to reduce cognitive decline in patients with amnestic mild cognitive impairment due to AD. We used two curated datasets from the ADNI study representing weakly and strongly prognostic baseline variables for the outcome to pre- or post-select baseline variables for use in hypothetical trials to which these pre- or post-selected variables were applied. We consider several baseline variable selection procedures including always adjusting for the baseline outcome only, adjusting for all 21 pre-specified candidate baseline variables, and adjusting for variables selected from a proposed cross-validated R^2 procedure ($CV - R^2$), the lasso or VSURF (random forest) procedures.

All of our estimators had similar and small bias, and the corresponding confidence intervals produced roughly 95% coverage of the marginal treatment effect.Adjusting for prognostic baseline variables selected from the CV-R^2, lasso and RF procedures, as well as including all candidate baseline variables, resulted in large reductions in the required sample size when compared to the unadjusted estimator (roughly 15 to 30% reduction). Selecting baseline variables using the lasso procedure resulted in adjusted marginal treatment effects with the largest precision gains. When baseline variables are not prognostic, the lasso resulted in approximately no loss of precision.

We recommend baseline variable adjustment within randomized trials where there are prognostic baseline variables. The baseline variable selection procedure should be pre-planned including when and how baseline variables are selected. Post-selecting baseline variables using the lasso procedure resulted in the largest precision gain when baseline variables are prognostic for the outcome and small loss of precision when baseline variables are not prognostic.

Thesis Committee

Elizabeth Colantuoni (Primary Advisor)
Senior Scientist
Department of Biostatistics
Johns Hopkins Bloomberg School of Public Health

Michael Rosemblum
Professor
Department of Biostatistics
Johns Hopkins Bloomberg School of Public Health

Acknowledgements

First and foremost, to my primary advisor, Elizabeth Colantuoni, I can't thank you enough for your guidance, encouragement, kindness, and patience from the start of this program. You have truly made me feel supported through this program, and helped me grow as a writer and researcher.

To my advisor, Michael Rosemblum, I thank you for the abundant advice on the analysis and construction of the paper, and the support you have provided me. It has truly been a joy to learn from both of you.

To Arnold Bakker, I thank you for your extensive knowledge in Alzheimer's Disease and Mild Cognitive Impairment, and your assistance in the construction of the paper.

To both Arnold Bakker and Michela Gallagher, I thank for your work and guidance on the HOPE4MCI trial and Alzheimer's Disease Neuroimaging Initiative study.

To my friends and family, I express my deepest gratitude for supporting me, uplifting me, and making me laugh. To my father, who stayed on the phone with me through every difficult moment.

To all the friends I made within the Biostatistics department and Johns Hopkins Bloomberg School of Public Health, thank you for the laughs, kindness, and friendships that uplifted me.

Lastly, to my partner, Zach, thank you for the unconditional love and support. Thank you for always trying to call and check in from 3,000 miles away. I couldn't have done it without you all!

Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at: http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf

ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: AbbVie, Alzheimer's Association; Alzheimer's Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen; Bristol-Myers Squibb Company; CereSpir, Inc.; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Lumosity; Lundbeck; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical Company; and Transition Therapeutics. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Disease Cooperative Study at the University of California, San Diego. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California. This work is solely the responsibility of the authors and does not represent the views of the above people and agencies.

Table of Contents

List of Tables

List of Figures

1 Introduction

We focus on the context of a randomized controlled trial (RCT) comparing a treatment to control with the goal of estimating the marginal treatment effect. While an unadjusted estimator of the marginal treatment effect is simple to implement and unbiased, the estimator disregards potential information in the baseline variables. Adjusting for baseline variables that are correlated with the outcome can improve, i.e. increase, the precision of the estimated marginal treatment effect (*1-8*).

Adjusting for baseline variables within RCTs has been a much discussed topic (*9-10*). Reviews of RCTs have found that only 24% to 34% of trials adjust for baseline variables in the analysis (*11-14*). In 2010, the Consolidated Standards of Reporting Trials (CONSORT) provided clarity surrounding study methods, like baseline variable adjustment, for RCTs (*15*). They suggested "...an adjusted analysis may be sensible, especially if one or more variables is thought to be prognostic", "the decision to adjust should not be determined by whether baseline differences are statistically significant" and "adjusted analyses should be specified in the study protocol" (p. 14).

Although an extensive literature exists on a multitude of estimators for the marginal treatment effect that adjust for baseline variables and the large sample properties of these estimators (*8, 16-20*), there has been less guidance on what procedure to use to select baseline variables for use with these estimators. Our work is motivated by the ongoing HOPE4MCI trial (*21*). The primary objective of the trial is to assess the efficacy of a drug, AGB101, compared to a placebo on slowing cognitive decline in patients with amnestic mild cognitive impairment (MCI) due to Alzheimer's disease (AD).

While planning the trial, the statistical analysis plan was developed including specification of which baseline variables are to be included in the estimation of the marginal treatment effect for the primary outcome. However, the question remains of

whether the selection procedure, i.e. pre-selecting baseline variables for adjustment, was optimal in the sense of maximizing the potential precision gain for the marginal treatment effect. As an alternative, baseline variables could be post-selected, i.e. identified during the analysis of the RCT data. The differences in these two approaches have not been evaluated to the best of our knowledge.

Further, regardless of when (pre or post) the baseline variables are selected, there are multiple baseline variable selection methods, but a lack of guidance as to which is preferred. A wide range of classical variable selection methods exist, such as stepwise, forward, and backward selection procedures. More recently there has been rapid development in modern machine learning based variable selection methods. Machine learning methods, such as lasso regression, have been used to estimate the marginal treatment effect (*23-28*). Bloniarz et al.(2016) evaluated properties of the baseline covariate adjusted marginal treatment effect derived under the randomization inference framework, which differs from the superpopulation inference framework used here. They showed that first using the cross-validated lasso (CV lasso) for variable selection followed by fitting a linear model with ordinary least squares regression for estimation, can substantially reduce the mean squared error (MSE) in estimating the marginal treatment effect. Wager et al. (2016), using superpopulation inference, proposed estimators of the marginal treatment effect using a CV lasso procedure and a CV random forest procedure. The approaches by Bloniarz et al.(2016) and Wager et al. (2016) resulted in substantial variance reductions compared to the unadjusted estimator of the marginal treatment effect, i.e. difference in sample means. All of these methods had similar confidence interval coverage probability.

Unlike the previously mentioned papers, Gagnon-Bartsch et al. (2020) make use observational data from a previous study to try to increase the precision in the estimated marginal treatment effect in a randomized trial, while guaranteeing unbiasedness. They propose the reLOOP procedure. The first step is to apply a

machine learning model, e.g. lasso regression or random forest, for the outcome given baseline variables using the observational data (referred to as the **re**mnant). the remnant is then used to predict the outcomes for each individual in the RCT as a function of their baseline variables. Then they generate a leave-one-out potential outcome (LOOP) prediction under treatment and control for each individual by fitting treatment-group specific models for the outcome as a function of baseline variables and the predicted outcome from the remnant, using the data from the N-1 remaining individuals (which leads to unbiasedness). These models can be fit by any method ranging from linear regression to random forests.

We use the context of the HOPE4MCI trial to systemically evaluate the impact of when and how the baseline variables are selected for covariate adjustment. In all cases, the selected baseline variables will be used in a linear regression model of the outcome given treatment and baseline variables that is fit with ordinary least squares to estimate the marginal treatment effect using the analysis of covariance (ANCOVA) method. The timing of the variable selection can be either pre-trial (called pre-selection) or based on the data accrued at the end of the trial (called post-selection). In all cases, the variable selection is based on evaluating how prognostic the baseline variables are for the outcome (and not on how imbalanced the baseline variables are across arms). The paper is organized as follows. In the next section, we provide notation and definitions followed by descriptions of several variable selection procedures. In sections 3 and 4, we describe and provide results from an extensive simulation study motivated by the HOPE4MCI trial. In the discussion, we summarize our findings and provide future areas of research.

2 Methods

2.1 Notation and Definitions

We assume a randomized controlled trial where we observe n independent participants, each with data vector (W_i, A_i, Y_i) from an unknown probability distribution P, where W_i is a $m \times 1$ column vector of baseline variables, A_i is the treatment arm indicator ($A_i \in 0, 1$ where 1= treatment and 0= control), and Y_i is a continuous valued outcome. We assume no missing values. We assume 1:1 randomization so that treatment and placebo was assigned by the binary treatment arm indicator, A, by taking a draw from a Bernoulli distribution with probability 1/2. Thus the treatment arm is assigned independently of the baseline variables. The select baseline variables, W, can be continuous, binary, or categorical variables. The treatment effect of interest is the average or marginal treatment effect:

$$\psi = E(Y|A = 1) - E(Y|A = 0).$$

We estimate ψ using the analysis of covariance (ANCOVA) estimator. The ANCOVA estimator of the marginal treatment effect adjusts for chance imbalance between the treatment arms in baseline variables. The ANCOVA estimator is the coefficient for treatment from the following linear regression model regressing the outcome Y on the treatment arm indicator A and main terms for baseline variables W:

$$E[Y|A, W] = \beta_0 + \beta_1 A + \beta_2 W_1 + ... + \beta_{m+1} W_m. \quad (1)$$

The estimate of ψ is $\widehat{\beta}_1$ obtained via ordinary least squares (OLS). Yang and Tsiatis (2001) showed that the ANCOVA estimator is consistent, even when the linear regression model is arbitrarily misspecified.

For the remainder of the paper, we use the ANCOVA estimator for the marginal

treatment effect due to its simplicity. However, there are alternatives to the this estimator that have favorable properties beyond consistency, e.g. Rotnitzky et al (2012) proposed a doubly robust estimator that is asymptotically guaranteed to be as precise or more precise than the unadjusted estimator. See (*8*) for a comparison of several additional estimators.

2.2 Pre/Post selection of variables for inclusion in ANCOVA estimator

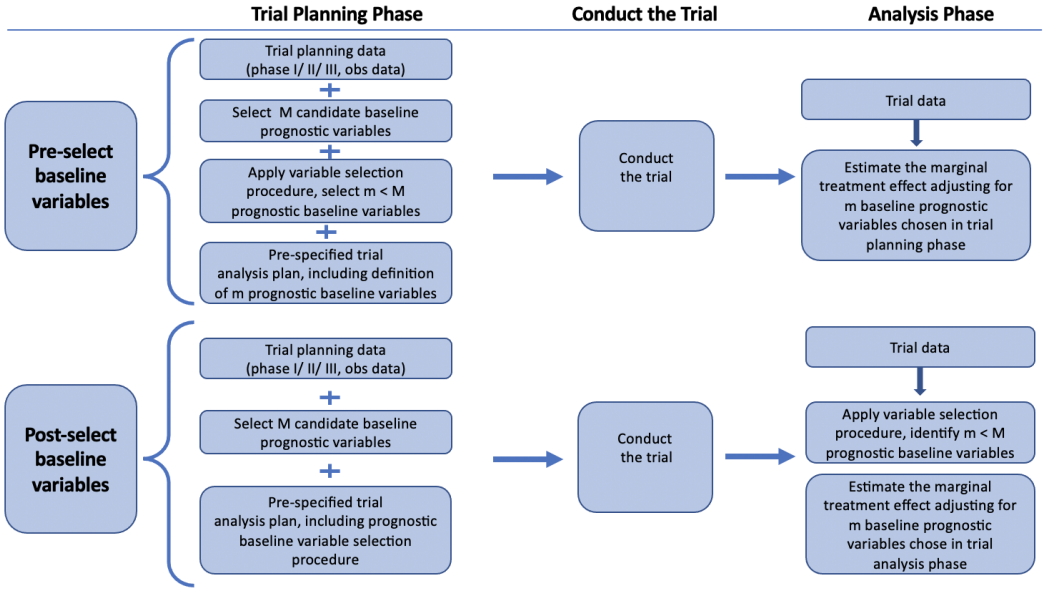

Figure 1: We can pre-select or post-select baseline variables. If variables are pre-selected, identify a set of potentially prognostic variables to choose from the available trial planning data and select m baseline variables during the planning phase. Using the trial data in the analysis phase, estimate the marginal treatment effect adjusting for the m prognostic baseline variables chosen in the trial planning phase. If variables are post-selected, identify a set of potentially prognostic variables to choose from the available trial planning data during the planning phase. Using the trial data in the analysis phase, select the m prognostic variables and estimate the marginal treatment effect adjusting for the m prognostic baseline variables chosen in the analysis stage.

Typically, the baseline variables to be included in the ANCOVA estimator are pre-selected, i.e., selected before the trial. We also consider here the case where they are post-selected, i.e., selected at the end of the trial based on data collected in the

trial. Figure 1 shows the difference between pre- and post-selection. In both cases, we assume a common set of M candidate baseline variables that may be selected for inclusion in the vector W used in the ANCOVA estimator. The candidate list of variables would typically be constructed based on clinical knowledge, previous data sets, and relevant existing literature.

When we pre-select W, it is based on a separate, pre-existing dataset (e.g., from an observational study or trial involving the same population) that contains both Y and the initial set of M candidate baseline variables. A variable selection procedure is applied to that previous dataset to select $m < M$ baseline variables that will be named in the analysis plan for the trial. At the conclusion of the trial, the marginal treatment effect is estimated by ANCOVA that adjusts for the pre-specified m variables in the ANCOVA estimator.

An alternative is to post-select the m baseline variables using the data in the trial itself. Here, while planning the trial, the set of M candidate baseline variables would be identified, again based on existing knowledge of the relationship between Y and the M candidate baseline variables, and the analysis plan would be developed, including details of how the trial data will be used to select m of the M candidate baseline variables for inclusion in the ANCOVA estimator. The number m of variables included in the ANCOVA estimator is not fixed for the case of post-selection, since it depends on the variables selection procedure's output (which, in turn, depends on the trial data).

2.3 Variable Selection Procedures

Regardless of the timing of the selection of the baseline variables, we use a variable selection procedure. In this manuscript, we consider three variable selection procedures defined below.

2.3.1 Cross-validated R^2

We refer to the first procedure as the cross-validated R^2 (CV-R^2) procedure. It involves a structured search over candidate subsets W of the M baseline variables, where additional baseline variables are included if they increase the estimated relative efficiency of the unadjusted estimator by a non-negligible amount.

For a candidate subset W of the M baseline variables, we estimate the additional variation in Y that can be explained by W using OLS regression, beyond what can be explained by A alone (*19*), called the R-squared. This estimate involves computing the sum of squared residuals for Y based on estimates of the study arm specific mean of Y, sm_1 and sm_0 for the treatment and control arms, respectively. Second, use ordinary least squares regression separately by study arm ($a \in 0, 1$) to obtain the model fit $Q^a(W, B^{(a)})$ for the linear regression model $E(Y = 1|A = a, W) = \beta_{0a} + \beta_{1a}W_1 + ... + \beta_{m,a}W_m$ where $B^{(a)}$ denotes the regression coefficients from the study arm specific regression model. To approximate the relative reduction in required sample size to achieve a desired power using the adjusted estimator (ANCOVA) compared to the unadjusted estimator for the marginal treatment effect (*38, 39*), we define the estimated R-squared as

$$\widetilde{R_n^2} = 1 - \frac{\sum_{i=1}^{n}(Y_i - Q^{A_i}(W, \widehat{B}^{A_i}))^2}{\sum_{i=1}^{n}(Y_i - sm_{A_i})^2}. \tag{2}$$

To avoid being overly optimistic, we use a modified estimate of R-squared that is similar to the above display except using a leave-one-out cross-validation (CV) procedure. The leave-one-out CV is implemented by fitting sm_a and $Q^a(W, B^a)$ on all the data except observation i. Then for each participant i, compute the squared difference between Y_i and the estimates of $Q^a(W_i, \widehat{B}^a)$ and $\widehat{sm}_a$ and replace these in the sums in the numerator and denominator, respectively, on the right side of (2). The corresponding cross-validated estimator of the R-squared (2) is denoted by $\widehat{R_n^2}$

instead of $\widetilde{R_n^2}$.

The CV-R^2 baseline variable selection procedure is the following:

Step 1: Compute an estimate of $\widehat{R_n^2}$ when adjusting for each of the M baseline variables one at a time. Rank these from highest (largest R-squared) to lowest (smallest R-squared).

Step 2: If the highest ranked variable has $\widehat{R_n^2} \leq \frac{1}{n}$, then set the baseline variable set to be empty, which is equivalent to using the unadjusted estimator.

Else, do a single pass over each candidate variable from highest to lowest rank to determine which variables get selected. To do this, first initialize the current variable set to be empty. Next, considering the candidate variables from highest to lowest rank, include each candidate variable whose addition to the current variable set increases the corresponding $\widehat{R_n^2}$ by at least $1/n$.

2.3.2 Lasso Variable Selection

Tibishirani (1996) proposed the least absolute shrinkage and selection operator (lasso) regression, a popular machine learning technique for model selection based on a set of candidate covariates. The goal of the lasso regression is to determine the most important predictors to include for a reduced, parsimonious model. The lasso regression includes a L_1 penalty on the sum of the absolute value of the coefficients in addition to the ordinary least squares (OLS) loss function (i.e. the sums of squared residuals). As a result, some coefficients may be forced to 0 and discarded from the model. Therefore, the variable selection process is embedded in the model due to the L_1 penalty. The lasso's properties under linear regression, including consistent coefficient estimates, have been established (*36*).

The geometry of the L_1 penalty of the lasso creates the parsimonious model, making the method popular in high-dimensional data analysis, especially when the

number of covariates are larger than the sample size. In such settings, overfitting can be present in linear regression based on OLS. Sometimes individuals conduct model selection based on how imbalanced covariates are after randomization, leading to incorrect inference (*32*). Lasso can mitigate these problems by performing variable selection and avoiding overfitting (*33*). In addition, Tibshirani (1996) showed that the lasso is more accurate and stable than traditional methods like the best subset selections when there are small to moderate number of moderate-sized effects and a large number of small effects. Hastie et al (2020) concluded that lasso gave better accuracy than forward stepwise selection and best subset selection in low signal-to-noise (SNR) range, and the relaxed lasso provided the highest accuracy in all SNR levels (*43*).

2.3.3 Random Forest

Another popular machine learning technique for variable selection and building prediction models is random forests (RF), first introduced by Breiman (2001). A RF is a collection of classification or regression decision trees modeled off a randomly selected training set and a subset of pre-specified predictor variables. Within each decision tree, binary recursive partitioning is used to partition the covariate space such that the mean of observations falling into a given partition defines the predicted value for the outcome. Outcome predictions are determined by the aggregated results from a large number of trees. Therefore, compared to a single decision tree, a RF typically provides higher accuracy and maintains an interpretable relationship between the outcome and predictors (*38*). A RF is able to perform variable selection by determining the statistical characteristics of the optimal predictors, including accuracy of prediction and variable importance, i.e. how often variables are included across the trees in the forest and the improvement in prediction accuracy when they are selected. In this way, similar to the lasso, a RF can determine the most important variables for a

parsimonious and efficient model.

VSURF is a type of RF where variable selection is conducted by a two-step algorithm (*39*). First, rank the variables in descending order by an averaged variable importance (VI) over roughly 50 decision trees. Eliminate the unimportant variables that do not exceed a certain variable importance threshold. Second, the algorithm selects two subsets of variables: interpretation and prediction variables. For interpretation, the algorithm computes a nested group of classification and regression trees and ends with a group of variables that are all highly correlated with the outcome. For prediction, a stepwise sequence of nested classification and regression trees are constructed using the variables chosen in the interpretation step. The prediction step leads to a smaller subset of variables that have minimal redundancy for prediction purposes. We considered the subset of variables chosen in the prediction step as the variables chosen by the VSURF algorithm.

3 Simulation Study

3.1 HOPE4MCI Trial

Our work is motivated by the HOPE4MCI clinical trial, the main objective of which is to evaluate the efficacy of a the experimental drug, AGB101, on slowing cognitive and functional impairment among patients with aMCI due to AD. Amnestic mild cognitive impairment (aMCI) is classified as a clinical condition defined by memory concerns and generally considered a transitional stage between normal aging and AD dementia. Studies of patients with aMCI have reported aberrant activation of the hippocampus, a structure critically important for episodic memory function (*40-42*). The observed hippocampal hyperactivity is now considered a characteristic feature of the aMCI stage of AD (*43*) and has been shown contribute to amyloid accumulation (*44-45*) and is correlated with disease progression (*46*). Treatment

with low dose Levetiracetam, an anti-seizure medication, has been shown to normalize hippocampal dysfunction and improve memory function in patients with aMCI (*42, 47*). The HOPE4MCI study is designed to further examine if low dose Levetiracetam can improve mild memory problems and slow progression of aMCI due to AD. The trial uses the Clinical Dementia Rating-Sum of Boxes (CDR-SB) score to quantify impairment in cognitive function, where scores range from 0 to 18 with higher scores indicating greater impairment. The primary outcome is the 18-month change in CDR-SB score. Enrolled patients were required to be 55 to 85 years old, have a baseline CDR-SB score of ≤ 2.5, and meet criteria for aMCI due to AD (*48*). The target sample size was 160 participants and the hypothesized treatment effect was a 30% reduction in the mean 18-month change in CDR-SB comparing the treatment to placebo arm, assuming the mean 18-month change in CDR-SB is 1.3 in the placebo arm.

3.2 Trial planning data

The Alzheimer's Disease Neuroimaging Initiative (ADNI) database was used to curate two datasets to aid in the design of the HOPE4MCI trial. The ADNI is a longitudinal multicenter cohort study with observational data from cognitively normal older adults, patients with MCI and patients with mild to moderate AD dementia. The longitudinal data includes biochemical, clinical, genetic and imaging biomarkers for the detection and prediction of AD progression. Both datasets were curated to match the inclusion and exclusion criteria and duration of the HOPE4MCI Trial, i.e. provide baseline to 18-month follow-up with CDR-SB scores as well as a set of potentially prognostic baseline variables. Specifically, at the time of selection, all participants with a consensus diagnosis of MCI, aged between 55-85 at baseline, a Global CDR score of 0.5, a Mini-Mental Status Exam Score ≥ 24, and a CSF amyloid concentration ≤ 192 pg/ml, for which 24 months of cognitive and neuroimaging data was available, were

identified. Key measures obtained from neuroimaging are used as secondary endpoints in the HOPE4MCI trial, but were not considered for this manuscript. The resulting dataset consisted of about 200 patients with MCI who met all criteria for entry into the HOPE4MCI study with 2-year longitudinal clinical and neuropsychological data available in the ADNI study. The available data was used to create two separate datasets. The first dataset uses the 6-month ADNI visit as the study baseline. The second dataset uses the baseline ADNI visit as the study baseline. The former dataset represents a somewhat later stage of progression, and we are interested in comparing the impact of different covariate-adjusted estimators using data generating distributions from both datasets.

For both datasets, the following are the M candidate baseline variables: age, gender, marriage status, divorce status, APOE4, CDR-SB score, Mini-Mental State Examination (MMSE) score, Logical Memory Delayed Recall Score, Modified Hachinski Total Score, Geriatric Depression Scale Total, Category Fluency (Animals) - Total Correct, Trail Making Test Part A - Time to Complete, Trails A Errors of Commission, Trail Making Test Part B - Time to complete, Trails B Errors of Commission, Trails B Errors of Omission, Alzheimer's Disease Assessment Scale- Cognition (ADAS Cog) 11 item score, ADAS Cog 13 Item score, Rey auditory verbal learning test (RAVLT) Delay, RAVLT Recognition, and Functional Activities Questionnaire (FAQ). Descriptive analysis of these two datasets revealed weaker correlations between the 18-month change in CDR-SB scores and the above baseline variables in the first compared to the second dataset. Below, we use the first and second ADNI datasets to represent the settings where Y and W are weakly or strongly correlated, respectively.

3.3 Data generating distributions

We first consider the case where the baseline variables are pre-selected, i.e., selected based only on a pre-trial data set. Below, we describe how we construct the following

three scenarios:

D1) Y and W are weakly correlated (as in dataset 1)

D2) Y and W are strongly correlated (as in dataset 2)

D21) Y and W are strongly correlated pre-trial (as in dataset 2) but weakly correlated in the trial itself (as in dataset 1).

To evaluate the performance of post-selecting the baseline variables, we considered only D1 and D2 (since in this case it's irrelevant what the pretrial data is).

In each scenario, we considered four simulation settings:

1) Baseline variables prognostic and no marginal treatment effect.

2) Baseline variables prognostic and positive marginal treatment effect.

3) Baseline variables not prognostic and no marginal treatment effect.

4) Baseline variables not prognostic and positive marginal treatment effect.

The positive marginal treatment effect is defined as a reduction in the 18-month change in CDR-SB scores by 30% in the treatment arm compared to the control arm.

For each scenario-setting pair, we generated 10,000 simulated trials, each with $n = 160$ participants, the target sample size of the HOPE4MCI trial. In all cases, we first re-sampled vectors (W, Y) (where here W is the full list of M candidate baseline variables) with replacement from the appropriate ADNI dataset (depending on D1, D2, D21). The reason is that we try to mimic the correlations from the corresponding ADNI data sets, to make the simulations realistic. Next the treatment arm indicator is assigned independent of the baseline variables, with probability 1/2 for treatment and placebo. The resulting data generating distribution corresponds to baseline variables being prognostic for the outcome and to no treatment effect (setting 1). We next describe modifications that we made to define settings 2-4, which include either/both of the following: (i) making the baseline variables not prognostic for (i.e., independent of) the outcome, and (ii) inducing a positive treatment effect.

For (i), nothing is changed for settings 1 and 2. In simulation settings 3 and 4,

we create a distribution where the baseline variables are not prognostic by randomly permuting the Y values among the trial participants. For (ii), in simulation settings 2 and 4, we induce a marginal treatment effect of 0.3 by replacing each treatment arm participant's value of Y by a 30% reduction defined as $Y - 0.3 * Y = 0.7Y$.

In order to evaluate the performance of the methods that pre-select baseline variables, prior to generating any simulated trials, we constructed a fixed, pre-trial (planning) dataset for every combination of scenario (D1,D2,D21) and setting (1-4). These datasets are based, roughly, on applying the aforementioned method to the corresponding ADNI dataset. Specifically, for scenario D1 we started with ADNI dataset 1; for D2 and D21 we started with ADNI dataset 2. Next, the treatment arm indicator was assigned independently of the baseline variables, with probability 1/2, to each participant. This corresponds to setting 1. To construct planning data sets for scenarios 2-4, we make modifications described next. For simulation settings 3 and 4, we randomly permuted the Y values among the participants. For simulation settings 2 and 4, we induced a marginal treatment effect of 0.3 by replacing each treatment arm participant's value of Y by a 30% reduction. Each scenario by setting combination was generated one time with the full ADNI dataset 1 or 2. These are considered fixed. Therefore, each baseline variable selection procedure for the pre-selection case needs to be run only once for each scenario by setting combination.

We next describe the estimators and confidence interval procedures that we applied to each simulated trial. For each simulated trial, we estimated the marginal treatment effect using the unadjusted estimator and 5 ANCOVA estimators adjusting for the following baseline variables, respectively:

A1) Baseline CDR-SB score only.

A2) All M candidate prognostic baseline variables.

A3) The m baseline variable selected via the CV-R^2 procedure.

A4) The m baseline variables selected via the lasso regression procedure.

A5) The m baseline variables selected via the RF procedure.

Confidence intervals were defined by: $\widehat{\beta_1} \pm 1.96 * se(\widehat{\beta_1})$, with $se(\widehat{\beta_1})$ denoting the estimated standard error returned by the corresponding OLS linear regression model fit for (1), computed as if the baseline variable set were fixed in advance (for all cases).

4 Simulation Results

4.1 Overall Findings

Tables 1-8 display the simulation study results. Regardless of when and which baseline variables are selected, all estimators have similar and small bias and produced roughly 95% coverage of the marginal treatment effect, with coverage ranging from 92.8% to 95.3%. Further, the simulation results are similar under no or positive marginal treatment effect.

We quantified the precision gain when using an adjusted estimator compared to the unadjusted estimator by estimating the reduction in required sample size when using the adjusted estimator compared to the unadjusted estimator with fixed power: 1 - 1/relative efficiency, where relative efficiency is MSE(unadjusted) / MSE(adjusted). When comparing the adjusted estimators, adjusting for only the baseline CDR-SB score resulted in the smallest gain in precision over using the unadjusted estimator when baseline variables are prognostic (reduction in required sample size ranging from -0.2% to 7.7%). Adjusting for all M candidate prognostic baseline variables resulted in the largest precision loss when baseline variables are not prognostic (sample size reduction ranging from -17.3% to -16.1%) and performed similarly to the three variable selection procedures when baseline variables are prognostic. As we highlight further below, in all scenarios with prognostic baseline variables, the lasso procedure resulted in the largest precision gains, and in all scenarios with the no prognostic baseline variables, the lasso procedure resulted in the smallest precision loss.

4.2 Pre-selecting baseline variables

4.2.1 Selected baseline variables

Table 1: Baseline variables that get pre-selected from weakly correlated data (dataset 1) with a positive marginal treatment effect.

Baseline Variable	Prognostic Variables			No Prognostic Variables		
	$CV\text{-}R^2$	Lasso	RF	$CV\text{-}R^2$	Lasso	RF
CDR-SB	☐	☐	☑	☐	☐	☐
Age	☐	☐	☐	☐	☐	☐
Female	☐	☑	☐	☐	☐	☐
Married	☐	☐	☐	☐	☐	☐
Divorced	☐	☑	☐	☐	☐	☐
APOE4	☑	☑	☐	☐	☐	☐
MMSE	☐	☑	☐	☐	☐	☐
Logical Memory Score	☑	☑	☐	☐	☐	☐
Modified Hashinski	☐	☐	☑	☐	☐	☐
Geriatric Depression Scale Total	☐	☐	☐	☐	☐	☐
Category Fluency (Animals)	☐	☑	☐	☐	☐	☐
Trail Making Test A	☐	☐	☑	☐	☐	☐
Trails A Errors of Commission	☑	☑	☑	☐	☐	☐
Trail Making Test B	☑	☑	☐	☑	☐	☐
Trails B Errors of Commission	☐	☐	☐	☐	☐	☐
Trails B Errors of Omission	☐	☑	☑	☐	☐	☑
ADAS Cog 11 item score	☐	☐	☑	☐	☐	☑
ADAS Cog 13 item score	☑	☑	☑	☐	☐	☑
RAVLT Delay	☐	☑	☐	☐	☐	☐
RAVLT Recognition	☐	☐	☐	☐	☐	☐
FAQ	☐	☐	☐	☐	☐	☐

For scenarios D1 and D2/D21, we mimicked the planning phase of the HOPE4MCI trial by applying our baseline variable selection procedures to the first and second ADNI datasets, respectively, representing scenarios where baseline variables are weakly or strongly prognostic for the outcome. Table 1 and Table 2 display the baseline variables that were selected using each selection procedure. Regardless of the scenario, the variable selection procedures select very similar baseline variables for simulations with no (results not shown) or positive marginal treatment effect (Table 1 and 2). The variable selection procedures differ in the total number of variables selected and

Table 2: Baseline variables are pre-selected from strongly correlated data (dataset 2) with a positive marginal treatment effect.

Baseline Variable	Prognostic Variables			No Prognostic Variables		
	CV-R^2	Lasso	RF	CV-R^2	Lasso	RF
CDR-SB	☑	☑	☐	☐	☐	☑
Age	☐	☐	☐	☐	☐	☐
Female	☐	☐	☐	☐	☐	☐
Married	☑	☑	☐	☐	☐	☐
Divorced	☐	☐	☐	☐	☐	☐
APOE4	☑	☑	☐	☐	☐	☐
MMSE	☐	☑	☑	☐	☐	☐
Logical Memory Score	☑	☑	☐	☐	☐	☐
Modified Hashinski	☐	☑	☐	☐	☐	☐
Geriatric Depression Scale Total	☐	☑	☐	☐	☐	☐
Category Fluency (Animals)	☐	☐	☐	☐	☐	☑
Trail Making Test A	☐	☐	☐	☑	☐	☐
Trails A Errors of Commission	☐	☑	☑	☐	☐	☐
Trail Making Test B	☑	☑	☐	☐	☐	☐
Trails B Errors of Commission	☐	☐	☐	☑	☐	☐
Trails B Errors of Omission	☐	☑	☐	☐	☐	☐
ADAS Cog 11 item score	☐	☑	☑	☐	☐	☑
ADAS Cog 13 item score	☑	☑	☑	☐	☐	☐
RAVLT Delay	☐	☑	☑	☐	☐	☐
RAVLT Recognition	☐	☐	☐	☐	☐	☐
FAQ	☑	☑	☐	☐	☐	☐

which variables were selected.

When pre-selecting variables under a positive marginal treatment effect and weakly prognostic baseline variables (dataset 1), the CV-R^2, lasso and RF procedures selected 5, 11 and 7 baseline variables, respectively (Table 1). The lasso procedure selected: female, divorced, APOE4, MMSE, Logical Memory Delayed Recall Score, Category Fluency (Animals) - Total Correct, Trails A Errors of Commission, Trial Making Test Part B - Time to complete, Trails B Errors of Omission, ADAS Cog 13 Item score, and RAVLT Delay. The CV-R^2 procedure selected 5 of the 11 variables that the lasso procedure selected (APOE4, Logical Memory Score, Trails A Errors of Commission, Trail Making Test B, and ADAS Cog 13 item score); where as only

3 variables were in common when comparing the lasso and RF procedures (Trail A Errors of Commission, Trails B Errors of Omission and ADAS COG 13 item score). The RF procedure selected baseline CDR-SB score, Modified Hashinski, Trail Making Test A and ADAS COG 11 item score, all which were not selected by the other two procedures. When pre-selecting variables under positive marginal treatment effect and no prognostic baseline variables, the CV-R^2, lasso and RF procedure choose 1, 0, and 3 baseline variables, respectively. The CV-R^2 procedure selected the Trail Making Test B and the RF procedure selected Trails B Errors of Omission and the ADAS COG 11 and ADAS COG 13.

When pre-selecting variables from dataset 2, a slightly different pattern emerged (Table 2). Under a positive marginal treatment effect with strongly prognostic baseline variables, the selection procedures ranked the same in terms of the number of baseline variables selected: lasso (14 variables), CV-R^2 (7 variables) and RF (5 variables). However, both the CV-R^2 and the RF procedures selected a subset of the variables selected by the lasso procedure, albeit a different subset. When the outcome was scrambled to create a scenario with no prognostic baseline variables, as before the lasso procedure selected no baseline variables and the CV-R^2 and RF procedures selected a similar number of baseline variables, 2 and 3, respectively.

4.2.2 Results of adjusting for pre-selected baseline variables

Tables 3 through 6 summarize the results of hypothetical trials that adjust for pre-selected baseline variables from D1 and D2, as described above. Adjusting for baseline variables selected from the CV-R^2, lasso, and RF procedures, as well as including all M candidate baseline variables, resulted in large reductions in required sample sizes when compared to the unadjusted estimator. The magnitude of the reduction in required sample size is a function of the strength of the correlation between the outcome and baseline variables, with greater reductions in required sample sizes when

Table 3: Baseline variables prognostic are pre-selected with no marginal treatment effect. Comparison of the bias, variance, mean squared error and 95% coverage probability for the marginal treatment effect based on 10,000 hypothetical trials for estimator. The $\widehat{R_n^2}$ is the relative reduction in required sample size when using an adjusted estimator compared to the unadjusted estimator of the marginal treatment effect.

		No Marginal Treatment Effect				
Scenario		Bias	Var[1]	MSE[2]	$\widehat{R_n^2}$[3]	Cov[4]
D1	Unadj	0.004	0.084	0.084	0.0	0.949
	Y_0	0.004	0.084	0.084	-0.2	0.949
	All Cov	0.001	0.072	0.072	14.7	0.947
	CV-R^2	0.002	0.070	0.070	16.3	0.948
	Lasso	0.003	0.070	0.070	16.6	0.944
	RF	0.002	0.072	0.072	14.0	0.949
D2	Unadj	-0.001	0.072	0.072	0.0	0.948
	Y_0	-0.001	0.066	0.066	7.7	0.949
	All Cov	-0.002	0.050	0.050	29.9	0.951
	CV-R^2	-0.001	0.049	0.049	31.6	0.953
	Lasso	-0.002	0.049	0.049	31.9	0.951
	RF	-0.002	0.060	0.060	16.4	0.951
D21	Unadj	0.004	0.084	0.084	0.0	0.949
	Y_0	0.004	0.084	0.084	-0.2	0.949
	All Cov	0.001	0.072	0.072	14.7	0.947
	CV-R^2	0.002	0.072	0.072	14.1	0.947
	Lasso	0.002	0.070	0.070	16.5	0.947
	RF	0.004	0.075	0.075	11.2	0.950

[1] Var corresponds to variance

[2] MSE corresponds to mean squared error= $variance + bias^2$

[3] $\widehat{R_n^2}$ corresponds to the relative reduction in required sample size comparing the adjusted estimator to the unadjusted estimator, $1 - \frac{MSE(adjusted)}{MSE(unadjusted)}$

[4] Cov corresponds to coverage probability

the trial data includes strongly prognostic baseline variables (dataset 2) compared to weakly prognostic baseline variables (dataset 1). Within each scenario, adjusting for baseline variables selected from the lasso procedure yields the greatest reduction in required sample size when compared to the unadjusted estimator (roughly 15% or 30% reduction in required sample size when there is weak or strong correlation between baseline variables and the outcome, respectively). For example, in simulations with

Table 4: Baseline variables prognostic are pre-selected with a marginal treatment effect. Comparison of the bias, variance, mean squared error and 95% coverage probability for the marginal treatment effect based on 10,000 hypothetical trials for estimator. The $\widehat{R_n^2}$ is the relative reduction in required sample size when using an adjusted estimator compared to the unadjusted estimator of the marginal treatment effect.

		Marginal Treatment Effect				
Scenario		Bias	Var[1]	MSE[2]	$\widehat{R_n^2}$[3]	Cov[4]
D1	Unadj	-0.001	0.062	0.062	0.0	0.949
	Y_0	-0.001	0.062	0.062	-0.2	0.952
	All Cov	0.005	0.054	0.054	13.3	0.945
	CV-R^2	0.002	0.052	0.052	16.0	0.946
	Lasso	0.004	0.052	0.052	17.0	0.946
	RF	0.003	0.055	0.055	12.3	0.945
D2	Unadj	0.005	0.055	0.055	0.0	0.944
	Y_0	0.005	0.051	0.051	7.1	0.944
	All Cov	0.011	0.040	0.040	27.4	0.941
	CV-R^2	0.006	0.039	0.039	28.8	0.942
	Lasso	0.010	0.038	0.038	29.7	0.942
	RF	0.007	0.046	0.046	15.5	0.943
D21	Unadj	-0.001	0.062	0.062	0.0	0.949
	Y_0	-0.001	0.062	0.062	-0.2	0.952
	All Cov	0.005	0.054	0.054	13.3	0.945
	CV-R^2	-0.001	0.053	0.053	14.7	0.947
	Lasso	0.003	0.053	0.053	15.5	0.944
	RF	0.002	0.056	0.056	10.7	0.944

[1] Var corresponds to variance

[2] MSE corresponds to mean squared error= $variance + bias^2$

[3] $\widehat{R_n^2}$ corresponds to the relative reduction in required sample size comparing the adjusted estimator to the unadjusted estimator, $1 - \frac{MSE(adjusted)}{MSE(unadjusted)}$

[4] Cov corresponds to coverage probability

no marginal treatment effect and strongly prognostic baseline variables (dataset 2), adjusting for the baseline variables pre-selected by the lasso procedure resulted in largest sample size reduction (31.9%), compared to 31.6% and 16.4% sample size reduction when adjusting for variables selected with CV-R^2 and RF, respectively. Including all M baseline variables resulted in a sample size reduction of 29.9%.

Further, when baseline variables are not prognostic, identifying baseline variables

Table 5: Baseline variables not prognostic are pre-selected with no marginal treatment effect. Comparison of the bias, variance, mean squared error and 95% coverage probability for the marginal treatment effect based on 10,000 hypothetical trials for estimator. The $\widehat{R_n^2}$ is the relative reduction in required sample size when using an adjusted estimator compared to the unadjusted estimator of the marginal treatment effect.

		No Marginal Treatment Effect				
Scenario		Bias	Var[1]	MSE[2]	$\widehat{R_n^2}$[3]	Cov[4]
D1	Unadj	-0.001	0.085	0.085	0.0	0.949
	Y_0	-0.002	0.085	0.085	-0.5	0.950
	All Cov	-0.001	0.099	0.099	-17.3	0.948
	CV-R^2	-0.001	0.085	0.085	-0.7	0.950
	Lasso	-0.001	0.085	0.085	0.0	0.949
	RF	-0.002	0.088	0.088	-4.3	0.950
D2	Unadj	-0.004	0.071	0.071	0.0	0.952
	Y_0	-0.004	0.072	0.072	-0.7	0.952
	All Cov	-0.002	0.083	0.083	-16.2	0.950
	CV-R^2	-0.004	0.073	0.073	-2.1	0.951
	Lasso	-0.004	0.072	0.072	-0.6	0.950
	RF	-0.004	0.073	0.073	-2.2	0.950
D21	Unadj	-0.001	0.085	0.085	0.0	0.949
	Y_0	-0.002	0.085	0.085	-0.5	0.950
	All Cov	-0.001	0.099	0.099	-17.3	0.948
	CV-R^2	-0.001	0.086	0.086	-2.4	0.949
	Lasso	-0.001	0.085	0.085	-0.8	0.948
	RF	-0.001	0.086	0.086	-2.4	0.950

[1] Var corresponds to variance
[2] MSE corresponds to mean squared error= $variance+bias^2$
[3] $\widehat{R_n^2}$ corresponds to the relative reduction in required sample size comparing the adjusted estimator to the unadjusted estimator, $1 - \frac{MSE(adjusted)}{MSE(unadjusted)}$

using the lasso procedure yielded the smallest loss of efficiency compared to the unadjusted estimator, a less than 1% increase in required sample size (Table 5, Table 6). The remaining adjusted estimators can be ranked from smallest to largest increase in required sample size when compared to the unadjusted estimator: Baseline Y (less than 1%), CV-R^2 (1 to 2%), RF (2 to 4%) and All Covariates (roughly 16%).

Table 6: Baseline variables not prognostic are pre-selected with a marginal treatment effect. Comparison of the bias, variance, mean squared error and 95% coverage probability for the marginal treatment effect based on 10,000 hypothetical trials for estimator. The $\widehat{R_n^2}$ is the relative reduction in required sample size when using an adjusted estimator compared to the unadjusted estimator of the marginal treatment effect.

		Marginal Treatment Effect				
Scenario		Bias	Var[1]	MSE[2]	$\widehat{R_n^2}$[3]	Cov[4]
D1	Unadj	-0.006	0.061	0.061	0.0	0.951
	Y_0	-0.006	0.061	0.061	-0.7	0.951
	All Cov	-0.006	0.071	0.071	-16.6	0.950
	CV-R^2	-0.006	0.061	0.061	-0.5	0.951
	Lasso	-0.006	0.061	0.061	0.0	0.951
	RF	-0.006	0.063	0.063	-4.3	0.950
D2	Unadj	-0.001	0.055	0.055	0.0	0.948
	Y_0	-0.001	0.055	0.055	-0.2	0.949
	All Cov	-0.001	0.064	0.064	-16.1	0.947
	CV-R^2	-0.001	0.056	0.056	-1.5	0.948
	Lasso	-0.001	0.055	0.055	0.0	0.948
	RF	-0.001	0.056	0.056	-2.4	0.948
D21	Unadj	-0.006	0.061	0.061	0.0	0.951
	Y_0	-0.006	0.061	0.061	-0.7	0.951
	All Cov	-0.006	0.071	0.071	-16.6	0.950
	CV-R^2	-0.006	0.062	0.062	-1.3	0.951
	Lasso	-0.006	0.061	0.061	0.0	0.951
	RF	-0.006	0.062	0.062	-2.6	0.952

[1] Var corresponds to variance
[2] MSE corresponds to mean squared error= $variance+bias^2$
[3] $\widehat{R_n^2}$ corresponds to the relative reduction in required sample size comparing the adjusted estimator to the unadjusted estimator, $1 - \frac{MSE(adjusted)}{MSE(unadjusted)}$

4.3 Post-selecting baseline variables

4.3.1 Selected baseline variables

For scenarios D1 and D2, we simulate the HOPE4MCI trial where baseline variables are selected for inclusion in estimation of the marginal treatment effect during the analysis, assuming the baseline variables are weakly (dataset 1) or strongly (dataset 2) prognostic for the outcome. Table 7 and Table 8 display the baseline variables that were chosen using each selection procedure. In both scenarios, the variable

Table 7: Percentage of the 10,000 simulated trials where each post-selected baseline variable is selected, assuming weakly prognostic baseline variables (dataset 1) with a positive marginal treatment effect.

	Prognostic Variables			No Prognostic Variables		
	CV-R^2	Lasso	RF	CV-R^2	Lasso	RF
Average Number of Variables	7.9	12.3	6.5	3.0	1.1	2.1
Baseline Variable						
CDR-SB	39	48	74	14	6	9
Age	22	50	3	15	6	2
Female	34	56	2	14	6	1
Married	26	45	4	14	5	2
Divorced	18	55	3	17	5	4
APOE4	76	78	21	15	6	11
MMSE	30	56	26	13	5	12
Logical Memory Score	41	68	17	14	6	3
Modified Hashinski Score	19	48	36	16	6	4
Geriatric Depression Scale Total	17	42	50	15	6	11
Category Fluency (Animals)	49	68	21	14	6	15
Trail Making Test A	20	42	24	14	6	2
Trails A Errors of Commission	59	83	91	17	6	16
Trail Making Test B	68	85	30	12	5	5
Trails B Errors of Commission	27	48	6	15	5	3
Trails B Errors of Omission	29	61	82	18	5	32
ADAS Cog 11 item score	23	34	92	10	4	36
ADAS Cog 13 Item score	80	86	26	7	3	15
RAVLT Delay	52	82	38	13	5	13
RAVLT Recognition	34	48	2	13	5	7
FAQ	24	45	14	15	6	1

selection procedures choose similar baseline variables for simulations with no or a positive treatment effect. The average number of variables selected over all 10,000 simulations and which variables selected in each simulation differ across the variable selection procedures.

When post-selecting baseline variables in scenario D1 with weakly prognostic baseline variables and a positive marginal treatment effect, the average number of baseline variables chosen over the 10,000 hypothetical trials was 7.9, 12.3, and 6.5 for the CV-R^2, lasso, and RF procedures, respectively (Table 7). The 12 most frequently selected variables in the lasso procedure were (in order of most to least frequently

Table 8: Percentage of the 10,000 simulated trials where each post-selected baseline variable is selected, assuming strongly prognostic baseline variables (dataset 2) with a positive marginal treatment effect.

	Prognostic Variables			No Prognostic Variables		
	CV-R^2	Lasso	RF	CV-R^2	Lasso	RF
Average Number of Variables	8.5	13.7	6.5	3.0	1.2	2.0
Baseline Variable						
CDR-SB	46	73	47	13	6	12
Age	19	48	1	15	6	2
Female	20	42	3	14	6	2
Married	52	73	0	13	6	2
Divorced	26	42	13	15	6	4
APOE4	78	88	12	14	6	9
MMSE	39	65	59	14	6	11
Logical Memory Score	84	96	4	14	6	4
Modified Hashinski Score	41	71	4	15	6	6
Geriatric Depression Scale Total	24	59	10	16	7	12
Category Fluency (Animals)	25	52	33	14	6	15
Trail Making Test A	16	41	2	14	6	3
Trails A Errors of Commission	36	74	87	17	6	16
Trail Making Test B	76	92	9	12	5	6
Trails B Errors of Commission	24	48	4	14	5	4
Trails B Errors of Omission	25	71	80	18	6	25
ADAS Cog 11 item score	26	73	94	11	4	29
ADAS Cog 13 Item score	77	61	8	8	4	13
RAVLT Delay	19	60	25	13	5	12
RAVLT Recognition	23	44	90	13	5	10
FAQ	73	95	15	14	6	1

selected): ADAS Cog 13 item score (86%), Trail Making Test B (85%), Trails A Errors of Commission (83%), RAVLT Delay (82%), APOE4 (78%), Category Fluency (Animals) (68%), Logical Memory Delayed Recall Score (68%), Trails B Errors of Omission (61%), Female (56%), MMSE (56%), Divorced (55%), and age (50%). The CV-R^2 procedure selected 7 of the 8 variables that the lasso selected (ADAS Cog 13 item score (80%), APOE4 (76%), Trail Making Test B (68%), Trail Making Errors of Commission (59%), RAVLT Delay (52%), Category Fluency (Animals) (49%), Logical Memory Delayed Recall Score (41%)); while the RF only had 3 variables in common with the lasso (Trails A Errors of Commission (91%), Trails B Errors of Omission

(83%), and RAVLT Delay (38%)). The RF also selected ADAS Cog 11 item score (92%), baseline CDR-SB (74%), and Geriatric Depression Scale Total (50%).

When simulating no prognostic baseline variables under post-selecting baseline variables with a positive marginal treatment effect, the average number of baseline variables the CV-R^2, lasso, and RF procedures choose over all the hypothetical trials is 3.0, 1.1, and 2.2, respectively. The most frequently selected variables in the CV-R^2 procedure were the Trails B Errors of Omission (17.8%), Trails A Errors of Commission (17%), and Geriatric Depression Scale Total (16%), while the lasso selected the Modified Hashinski Total Score (6%), and the RF chose ADAS Cog item 11 score (36.3%) and Trails B Errors of Omission (33%).

Compared to post-selecting variables from scenario D1 with a positive marginal treatment effect and weakly prognostic baseline variables, when post-selecting variables in scenario D2 with a positive marginal treatment effect and strongly prognostic baseline variables (Table 8), the CV-R^2 and lasso variable selection procedures select 1 additional variable on average, and the RF selects roughly the same number of variables on average. The CV-R^2 and RF procedures selected a subset of the most frequently selected variables by the lasso, albeit a different subset. When the outcome was scrambled to create a scenario with no prognostic baseline variables, as expected the average number of baseline variables chosen is the same as in pre-selecting scenario D1 with no prognostic baseline variables.

4.3.2 Results of adjusting for post-selected baseline variables

Table 9 and Table 10 summarize the results of the hypothetical trials for scenarios D1 and D2, where prognostic baseline variables are post-selected. Similar to adjusting for pre-selected baseline variables, adjusting for post-selected baseline variables using the CV-R^2, lasso, and RF procedures, as well as including all the M covariates, produce large gains in efficiency when compared to the unadjusted estimator. The

Table 9: Baseline variables prognostic are post-selected with no marginal treatment effect. Comparison of the bias, variance, mean squared error and 95% coverage probability for the marginal treatment effect based on 10,000 hypothetical trials for estimator. The $\widehat{R_n^2}$ is the relative reduction in required sample size when using an adjusted estimator compared to the unadjusted estimator of the marginal treatment effect.

		No Marginal Treatment Effect				
Scenario		Bias	Var[1]	MSE[2]	$\widehat{R_n^2}$[3]	Cov[4]
D1	Unadj	0.004	0.084	0.084	0.0	
	Y_0	0.004	0.084	0.084	-0.2	0.949
	All Cov	0.001	0.072	0.072	14.7	0.947
	CV-R^2	0.003	0.073	0.073	13.6	0.933
	Lasso	0.002	0.066	0.066	21.2	0.949
	RF	0.002	0.073	0.073	13.6	0.948
D2	Unadj	-0.001	0.072	0.072	0.0	0.948
	Y_0	-0.001	0.066	0.066	7.7	0.949
	All Cov	-0.002	0.050	0.050	29.9	0.951
	CV-R^2	-0.001	0.051	0.051	28.7	0.940
	Lasso	-0.002	0.047	0.047	34.8	0.951
	RF	-0.002	0.056	0.056	21.4	0.952

[1] Var corresponds to variance

[2] MSE corresponds to mean squared error= $variance + bias^2$

[3] $\widehat{R_n^2}$ corresponds to the relative reduction in required sample size comparing the adjusted estimator to the unadjusted estimator, $\frac{1-MSE(adjusted)}{MSE(unadjusted)}$

[4] Cov corresponds to coverage probability

lasso procedure yields the greatest efficiency gains when compared to the unadjusted estimator (about 20% or 31-35% reduction in sample size when there is weak or strong correlation between baseline variables and the outcome, respectively). The CV-R^2 procedure and including all the M covariates result in similar sample size reduction of about 13-15% and 27-30%, and adjusting for variables selected by the RF procedure yields the smallest sample size reduction of roughly 13% and 19-21% when post-selected from dataset 1 and dataset 2, respectively.

When post-selecting baseline variables with no prognostic baseline variables (Table 11, Table 12), adjusting for baseline variables selected by the lasso procedure results in no greater than a 0.2% increase in required sample size. The RF procedure resulted in

Table 10: Baseline variables prognostic are post-selected with a marginal treatment effect. Comparison of the bias, variance, mean squared error and 95% coverage probability for the marginal treatment effect based on 10,000 hypothetical trials for estimator. The $\widehat{R_n^2}$ is the relative reduction in required sample size when using an adjusted estimator compared to the unadjusted estimator of the marginal treatment effect.

		Marginal Treatment Effect				
Scenario		Bias	Var[1]	MSE[2]	$\widehat{R_n^2}$[3]	Cov[4]
D1	Unadj	-0.001	0.062	0.062	0.0	0.949
	Y_0	-0.001	0.062	0.062	-0.2	0.952
	All Cov	0.005	0.054	0.054	13.3	0.945
	CV-R^2	0.004	0.055	0.055	12.5	0.928
	Lasso	0.022	0.050	0.050	19.4	0.941
	RF	0.005	0.054	0.055	12.5	0.945
D2	Unadj	0.005	0.055	0.055	0.0	0.944
	Y_0	0.005	0.051	0.051	7.1	0.944
	All Cov	0.011	0.040	0.040	27.4	0.941
	CV-R^2	0.008	0.040	0.040	26.5	0.929
	Lasso	0.028	0.037	0.038	30.8	0.940
	RF	0.010	0.044	0.045	18.6	0.942

[1] Var corresponds to variance

[2] MSE corresponds to mean squared error= $variance + bias^2$

[3] $\widehat{R_n^2}$ corresponds to the relative reduction in required sample size comparing the adjusted estimator to the unadjusted estimator, $\frac{1-MSE(adjusted)}{MSE(unadjusted)}$

[4] Cov corresponds to coverage probability

roughly a 1-2% increase in required sample size while the CV-R^2 procedure resulted in roughly a 7-8% increase in required sample size. Including all the M covariates yielded the greatest loss of precision, i.e. increase in required sample size of roughly 16-17%.

4.4 Comparing the timing of baseline variable selection

The results of our simulations were qualitatively similar when comparing the performance of the baseline variable adjusted marginal treatment effects to the unadjusted estimator when baseline variables are pre- or post-selected. There was no inflation in Type I error due to selecting baseline variables during the analysis compared to prior to

Table 11: Baseline variables not prognostic are post-selected with no marginal treatment effect. Comparison of the bias, variance, mean squared error and 95% coverage probability for the marginal treatment effect based on 10,000 hypothetical trials for estimator. The $\widehat{R_n^2}$ is the relative reduction in required sample size when using an adjusted estimator compared to the unadjusted estimator of the marginal treatment effect.

		No Marginal Treatment Effect				
Scenario		Bias	Var[1]	MSE[2]	$\widehat{R_n^2}$[3]	Cov[4]
D1	Unadj	-0.001	0.085	0.085	0.0	0.949
	Y_0	-0.002	0.085	0.085	-0.5	0.950
	All Cov	-0.001	0.099	0.099	-17.3	0.948
	CV-R^2	0.000	0.091	0.091	-7.2	0.937
	Lasso	-0.001	0.084	0.084	1.1	0.948
	RF	-0.001	0.086	0.086	-1.4	0.950
D2	Unadj	-0.004	0.071	0.071	0.0	0.952
	Y_0	-0.004	0.072	0.072	-0.7	0.952
	All Cov	-0.002	0.083	0.083	-16.2	0.950
	CV-R^2	-0.003	0.077	0.077	-8.0	0.938
	Lasso	-0.004	0.070	0.070	1.0	0.951
	RF	-0.004	0.072	0.072	-0.8	0.952

[1] Var corresponds to variance
[2] MSE corresponds to mean squared error= $variance+bias^2$
[3] $\widehat{R_n^2}$ corresponds to the relative reduction in required sample size comparing the adjusted estimator to the unadjusted estimator, $1 - \frac{MSE(adjusted)}{MSE(unadjusted)}$
[4] Cov corresponds to coverage probability

the trial. When using the CV-R^2 procedure, there was a roughly 3% larger gain in precision when baseline variables are prognostic and a roughly 5% smaller precision loss when baseline variables are not prognostic, when pre- compared to post-selecting baseline variables. When using the lasso procedure, there was a roughly 3% greater precision gain when selecting variables during the analysis compared to selecting variables prior to the trial. There is no difference between the precision gains when comparing the timing of the baseline variable selection using the RF procedure. Overall, the differences mentioned are small and there does not seem to be a clear advantage between pre-selecting and post-selecting baseline variables.

Table 12: Baseline variables not prognostic are post-selected with a marginal treatment effect. Comparison of the bias, variance, mean squared error and 95% coverage probability for the marginal treatment effect based on 10,000 hypothetical trials for estimator. The $\widehat{R_n^2}$ is the relative reduction in required sample size when using an adjusted estimator compared to the unadjusted estimator of the marginal treatment effect.

		Marginal Treatment Effect				
Scenario		Bias	Var[1]	MSE[2]	$\widehat{R_n^2}$[3]	Cov[4]
D1	Unadj	-0.006	0.061	0.061	0.0	0.951
	Y_0	-0.006	0.061	0.061	-0.7	0.951
	All Cov	-0.006	0.071	0.071	-16.6	0.950
	CV-R^2	-0.006	0.065	0.065	-6.7	0.939
	Lasso	0.002	0.060	0.060	1.1	0.950
	RF	-0.006	0.062	0.062	-1.8	0.950
D2	Unadj	-0.001	0.055	0.055	0.0	0.948
	Y_0	-0.001	0.055	0.055	-0.2	0.949
	All Cov	-0.001	0.064	0.064	-16.1	0.947
	CV-R^2	-0.001	0.059	0.059	-7.7	0.934
	Lasso	0.011	0.055	0.055	-0.2	0.945
	RF	-0.000	0.056	0.056	-1.5	0.947

[1] Var corresponds to variance
[2] MSE corresponds to mean squared error= $variance+bias^2$
[3] $\widehat{R_n^2}$ corresponds to the relative reduction in required sample size comparing the adjusted estimator to the unadjusted estimator, $1 - \frac{MSE(adjusted)}{MSE(unadjusted)}$
[4] Cov corresponds to coverage probability

5 Discussion

Based on the results of our simulation study, adjusting for baseline variables selected using the lasso procedure resulted in the largest precision gains for the marginal treatment effect, when baseline variables were prognostic for the outcome. When baseline variables were not prognostic for the outcome, the lasso procedure identified the fewest baseline variables for inclusion in the adjusted estimator resulting in the smallest loss in precision for the marginal treatment effect.

When considering a variable selection procedure, it is important to weigh both the advantages and disadvantages of the procedure, including complexity of implementation, computation time, and properties of the adjusted marginal treatment effect estimator.

Consistent with our findings, Tian et al. (2012) utilized the CV lasso procedure to select variables to include in adjusted estimators of the marginal treatment effect and found increased precision with superior computational efficiency compared to other methods including forward subset selection. In a comparison of RF and other tree-based and regression-based variable selection methods, Speiser et al. (2019) and Sanchez-Pineto et al. (2018) both noted VSURF was among the methods that selected the fewest variables (i.e. the most parsimonious model) and had the highest computation times due to it's stepwise procedure. While this was consistent with our findings, parsimony was also a possible reason why selecting baseline variables using VSURF did not result in as large a precision gain when compared to the lasso procedure. In simulations with prognostic baseline variables, the lasso regression picked roughly 12 variables on average compared to 8 for the VSURF procedure.

In addition to the choice of which baseline variable selection procedure to use, trialists must also decide when to select the baseline variables for use in an adjusted marginal treatment effect estimator: while planning the trial (pre-select variables) or during the analysis of the trial data (post-select variables). The key difference in the two approaches is that the first approach uses data available at the time of planning the trial to *a priori* specify baseline variables to include in the analysis of the trial; whereas the second approach, uses the data from the trial itself to both select baseline variables and estimate the marginal treatment effect. One potential drawback of the first approach is that the strength of associations between the outcome and baseline variables observed in the data used to plan the trial may be different then those observed in the trial data. We demonstrated this concern in simulation scenarios D2 and D21 where baseline variables were pre-selected from the second ADNI dataset, with strongly prognostic baseline variables, during the planning phase and then applied to subsequent trials where the baseline variables had similar prognostic power (scenario D2) vs. weak prognostic power (scenario D21). When the correlation in the

planning and trial data didn't match (D21), the precision gains were reduced by as much as 50% (similar to that achieved by pre- or post-selection under D1).

Some may interpret both the selection of baseline variables and estimation of the marginal treatment effect during the analysis of the trial as an opening for "gaming the system". This may be addressed by requiring that procedures are completely pre-specified in the study protocol and statistical analysis plan and subsequently followed. It is still important, however, to prove results on the asymptotics of the estimators (and confidence interval procedures) that use machine learning for variable selection.

One final consideration is the target sample size for a future trial. Should trialists assume efficiency gains from covariate adjustment and set the sample size accordingly? As mentioned previously, if one were to assume efficiency gains from an adjusted estimator and set the sample size based on this assumption, it is not guaranteed that the strength of the correlation between the baseline variables and the outcome will be similar in the trial data. If one were to assume no efficiency gains from an adjusted estimator, then precision gains can translate into improved power if they occur. However, one would not get any sample size reduction under the null hypothesis in this case. An alternative would be to use information-monitoring to set the sample size and/or trial duration; in this way, the observed correlations in the trial are used to estimate the variance (1 divided by the information). That way, more strongly prognostic baseline variables lead to faster information accrual and earlier analyses, i.e., shorter trials even under the null hypothesis of no treatment effect.

Our study has several limitations. In all simulation settings, we re-sampled (W, Y) with replacement to preserve the correlation patterns observed within the two ADNI datasets. There are certainly many alternative approaches for creating a set of W prognostic for Y. Given that we curated the ADNI datasets to mimic patient inclusion criteria for the HOPE4MCI trial, we feel that our simulations reflect potential efficiency gains within a real trial setting as opposed to a setting that is

generated for mathematical convenience. The foundation of our simulation study was data from the ADNI study, a longitudinal cohort. Although the two datasets we evaluated were curated to represent patients that may enroll within the HOPE4MCI trial, our results may differ if we had conducted our simulation study using data from previously completed AD trials, curated to match the HOPE4MCI inclusion/exclusion criteria. Next, when pre-selecting variables, the selection procedures were applied to the entire ADNI dataset. This gives an advantage for the pre-selection compared to post-selected algorithms that chose variables based on a sample of 160 participants from the entire population. Therefore, results from our pre-selection procedures may be overly optimistic. To avoid this, we could pre-select baseline variables using a bootstrap sample from the ADNI datasets. Further, we calculated all confidence intervals for the coverage probabilities with the standard error from the ANCOVA estimator, as if the selected baseline variables had been pre-specified in advance. Though this generally performed well in our simulations, for post-selecting variables, this method does not account for the variation in the baseline variable selection procedures. One could alternatively apply a non-parametric bootstrap procedure when post-selecting baseline variables.

Lastly, within the two machine learning variable selection procedures (lasso and random forest), we used the default settings specified in their respective R packages. Our results may be different if we first optimized these tuning parameters, e.g. adjusting the number of variables randomly sampled as candidates at each split in a RF or the tuning parameter of a lasso. Highly customizable algorithms could provide additional flexibility that could further improve precision of the marginal treatment effect and was not tested in our simulations. We have demonstrated the potential for substantial precision gains for the estimates of marginal treatment effects in AD trials Trialists should pre-specify the pool of candidate variables, selection procedures, and adjusted marginal treatment effect estimator.

Bibliography

[1] Hauck WW, Anderson S, Marcus SM. Should we adjust for covariates in nonlinear regression analyses of randomized trials? Control Clin Trials. 1998; 19(3):249–56.

[2] Ford I, Norrie J. The role of covariates in estimating treatment effects and risk in long-term clinical trials. Stat Med. 2002;21(19):2899–908.

[3] Ciolino JD, Martin RH, Zhao W, Jauch EC, Hill MD, Palesch YY. Covariate imbalance and adjustment for logistic regression analysis of clinical trial data. J Biopharm Stat. 2013;23(6):1383–402.

[4] Ciolino JD, Renee HM, Zhao W, Jauch EC, Hill MD, Palesch YY. Continuous covariate imbalance and conditional power for clinical trial interim analyses. Contemp Clin Trials. 2014;38(1):9–18.

[5] Hernández AV, Steyerberg EW, Habbema JDF. Covariate adjustment in randomized controlled trials with dichotomous outcomes increases statistical power and reduces sample size requirements. J Clin Epidemiol. 2004;57(5):454–60.

[6] Wang B, Ogburn EL, Rosenblum M. Analysis of covariance in randomized trials: More precision and valid confidence intervals, without model assumptions. Biometrics. 2019;75:1391–1400. https://doi.org/10.1111/biom.13062

[7] Díaz, I., Colantuoni, E., Hanley, D.F. and Rosenblum, M. (2018). Improved precision in the analysis of randomized trials with survival outcomes, without assuming proportional hazards. Lifetime Data Analysis, 1–30. https://doi.org/10.1007/s10985-018-9428-5.

[8] Colantuoni E, Rosenblum M. Leveraging prognostic baseline variables to gain precision in randomized trials. Stat Med. 2015 Aug 15;34(18):2602-17. doi:

10.1002/sim.6507. Epub 2015 Apr 14. Erratum in: Stat Med. 2017 Nov 30;36(27):4419. PMID: 25872751; PMCID: PMC5018399.

[9] Assmann SF, Pocock SJ, Enos LE, Kasten LE. Subgroup analysis and other (mis)uses of baseline data in clinical trials. Lancet. 2000; 355:1064–1069. [PubMed: 10744093]

[10] Pocock SJ, Assmann SE, Enos LE, Kasten LE. Subgroup analysis, covariate adjustment and baseline comparisons in clinical trial reporting: current practice and problems. Statistics in Medicine. 2002; 21(19):2917–2930. [PubMed: 12325108]

[11] Austin PC, Manca A, Zwarenstein M, Juurlink DN, Stanbrook MB: A substantial and confusing variation exists in handling of baseline covariates in randomized controlled trials: a review of trials published in leading medical journals. J Clin Epidemiol 2010, 63(2):142–153.

[12] Hernandez AV, Steyerberg EW, Taylor GS, Marmarou A, Habbema JD, Maas AI: Subgroup analysis and covariate adjustment in randomized clinical trials of traumatic brain injury: a systematic review. Neurosurgery 2005, 57(6):1244–1253. Discussion, 1253.

[13] Yu LM, Chan AW, Hopewell S, Deeks JJ, Altman DG: Reporting on covariate adjustment in randomised controlled trials before and after revision of the 2001 CONSORT statement: a literature review. Trials 2010, 11:59.

[14] Saquib N, Saquib J, Ioannidis JP: Practices and impact of primary outcome adjustment in randomized controlled trials: meta-epidemiologic study. BMJ 2013, 347:f4313.

[15] Moher D, Hopewell S, Schulz KF, Montori V, Gøtzsche PC, Devereaux P, Elbourne D, Egger M, Altman DG. CONSORT 2010 explanation and

elaboration: updated guidelines for reporting parallel group randomised trials. J Clin Epidemiol. 2010;63(8):e1–e37.

[16] Tan Z. Bounded, efficient and doubly robust estimating equations for marginal and nested structural models. Biometrika. 2010; 97:661–682.

[17] Rotnitzky A, Lei Q, Sued M, Robins JM. Improved double-robust estimation in missing data and causal inference models. Biometrika. 2012; 99(2):439–456. [PubMed: 23843666]

[18] Gruber S, van der Laan MJ. Targeted minimum loss based estimator that outperforms a given estimator. The International Journal of Biostatistics. 2012; 8(1) Article 11.

[19] Rubin D, van der Laan MJ. Empirical efficiency maximization: improved locally efficient covariate adjustment in randomized experiments and survival analysis. International Journal of Biostatistics. 2008; 4(1) Article 5.

[20] Moore K, van der Laan MJ. Covariate adjustment in randomized trials with binary outcomes: targeted maximum likelihood estimation. Statistics in Medicine. 2009; 28(1):39–64. [PubMed: 18985634]

[21] Mohs, R., Rosenzweig-Lipson, S., Barton, R. (2019, January 15 -). Study of AGB101 in Mild Cognitive Impairment Due to Alzheimer's Disease (HOPE4MCI). Identifier NCT03486938. https://clinicaltrials.gov/ct2/show/NCT03486938

[22] Miratrix, L. W., Sekhon, J. S., & Yu, B. (2013). Adjusting treatment effect estimates by post-stratification in randomized experiments. Journal of the Royal Statistical Society, Series B, 75, 369–396.

[23] Edward Wu and Johann A. Gagnon-Bartsch. The LOOP estimator: Adjusting

for covariates in randomized experiments. Evaluation Review, 42(4):458–488, 2018.

[24] Mark J Van der Laan and Sherri Rose. Targeted learning: causal inference for observational and experimental data. Springer Science Business Media, 2011.

[25] Bloniarz, A., Liu, H., Zhang, C., Sekhon, J. and Yu, B. (2016). Lasso adjustments of treatment effect estimates in randomized experiments. Proceedings of the National Academy of Sciences of the United States of America, 113, 7383–7390.

[26] Stefan Wager, Wenfei Du, Jonathan Taylor, and Robert J Tibshirani. High-dimensional regression adjustments in randomized experiments. Proceedings of the National Academy of Sciences, 113(45):12673–12678, 2016.

[27] Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. Double/debiased machine learning for treatment and structural parameters. The Econometrics Journal, 21(1):C1–C68, 2018.

[28] Gagnon-Bartsch, J.A., Sales, A., Wu, E., Anthony, F., Botelho, Miratrix, L., Heffernan, N. (2020). Precise Unbiased Estimation in Randomized Experiments using Auxiliary Observational Data.

[29] Yang L, Tsiatis AA. Efficiency study of estimators for a treatment effect in a pretest–posttest trial. The American Statistician. 2001; 55:314–321.

[30] Borm, G., Fransen, J. and Lemmens, W. (2007). A simple sample size formula for analysis of covariance in randomized clinical trials. Journal of Clinical Epidemiology, 60, 1234–1238.

[31] Rubin, D., and van der Laan, M. (2008). Covariate adjustment for the intention-to-treat parameter with empirical efficiency maximization. U.C.

Berkeley Division of Biostatistics Working Paper Series. Working Paper 229, https://biostats.bepress.com/ ucbbiostat/paper229.

[32] Permutt T (1990) Testing for imbalance of covariates in controlled experiments. Stat Med 9(12):1455–1462.

[33] Tibshirani R. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society. Series B (Methodological) 1996; 58(1):267–288.

[34] Trevor Hastie. Robert Tibshirani. Ryan Tibshirani. "Best Subset, Forward Stepwise or Lasso? Analysis and Recommendations Based on Extensive Comparisons." Statist. Sci. 35 (4) 579 - 592, November 2020. https://doi.org/10.1214/19-STS73

[35] Efron, B., Hastie, T., Johnstone, I., Tibshirani, R., 2004. Least angle regression. Ann. Statist. 32 (2), 407–499.

[36] Bühlmann P, Van De Geer S (2011) Statistics for High-Dimensional Data: Methods, Theory and Applications (Springer Science and Business Media, Berlin).

[37] L. Breiman. RFs. Machine Learning, 45(1):5–32, 2001. [p19, 20]

[38] Speiser JL, Durkalski VL, & Lee WM (2015). RF classification of etiologies for an orphan disease. Statistics in medicine, 34, 887–899. [PubMed: 25366667]

[39] Genuer, Robin Poggi, Jean-Michel Tuleau-Malot, Christine. (2015). VSURF: An R package for variable selection using RFs. The R Journal. 7. 10.32614/RJ-2015-018.

[40] Dickerson, B.C., Salat, D.H., Bates, J.F., Atiya, M., Killiany, R.J., Greve, D.N., Dale, A.M., Stern, C.E., Blacker, D., Albert, M.S., et al. (2004). Medial temporal lobe function and structure in mild cognitive impairment. Ann. Neurol. 56, 27-35.

[41] Yassa, M.A., Stark, S.M., Bakker, A., Albert, M.S., Gallagher, M., and Stark, C.E. (2010). High-resolution structural and functional MRI of hippocampal CA3 and dentate gyrus in patients with amnestic Mild Cognitive Impairment. Neuroimage 51, 1242-52.

[42] Bakker A., Krauss G.L., Albert M.S. et al. Reduction of hippocampal hyperactivity improves cognition in amnestic mild cognitive impairment. Neuron. 2012;74(3):467-474.

[43] Ewers M., Sperling R.A., Klunk W.E., Weiner M.W., Hampel H. Neuroimaging markers for the prediction and early diagnosis of Alzheimer's disease dementia. Trends in Neurosciences. 2011;34(8):430-442.

[44] Bero A.W., Yan P., Hoon Roh J., et al. Neuronal activity regulates the regional vulnerability to amyloid-deposition. Nat Neurosci. 2011;14(6):750-756.

[45] Leal S.L.,Landau S.M., Bell R.K., Jagust W.J. Hippocampal activation is associated with longitudinal amyloid accumulation and cognitive decline. eLife. 2017;6(Mci):1-15.

[46] Huijbers W., Schultz A.P., Papp K.V., et al. Tau Accumulation in clinically normal older adults is associated with increases in hippocampal fMRI activity. Alzheimers Dement. 2017;13(7):P1215.

[47] Bakker A., Albert M.S., Krauss G.,Speck CL., Gallagher M. Response of the medial temporal lobe network in amnestic mild cognitive impairment to therapeutic intervention assessed by fMRI and memory task performance. Neuroimage Clin. 2015;7:688-698.

[48] Albert MS, DeKosky ST, Dickson D, Dubois B, Feldman HH, Fox NC, Gamst A, Holtzman DM, Jagust WJ, Petersen RC, Snyder PJ, Carrillo MC, Thies B,

Phelps CH. The diagnosis of mild cognitive impairment due to Alzheimer's disease: recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. Alzheimers Dement. 2011 May;7(3):270-9. doi: 10.1016/j.jalz.2011.03.008. Epub 2011 Apr 21. PMID: 21514249; PMCID: PMC3312027.

[49] Sanchez-Pinto LN, Venable LR, Fahrenbach J, Churpek MM (2018). Comparison of variable selection methods for clinical predictive modeling. International journal of medical informatics, 116, 10–17. [PubMed: 29887230]

[50] Tian L, Cai T, Zhao L, Wei LJ. On the covariate-adjusted estimation for an overall treatment difference with data from a randomized comparative clinical trial. Biostatistics. 2012; 13(2):256– 273. [PubMed: 22294672]